Image and Vision Computing 65 (2017) 66–75

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Video-based emotion recognition in the wild using deep transfer learning and score fusion☆

Heysem Kaya[a,*], Furkan Gürpınar[b], Albert Ali Salah[b]

[a]Department of Computer Engineering, Çorlu Faculty of Engineering, Namık Kemal University, 59860 Çorlu, Tekirdağ, Turkey
[b]Department of Computer Engineering, Boğaziçi University, 34342 Bebek, İstanbul, Turkey

ARTICLE INFO

Article history:
Received 16 April 2016
Received in revised form 21 January 2017
Accepted 26 January 2017
Available online 4 February 2017

Keywords:
EmotiW
Emotion recognition in the wild
Multimodal fusion
Convolutional neural networks
Kernel extreme learning machine
Partial least squares

ABSTRACT

Multimodal recognition of affective states is a difficult problem, unless the recording conditions are carefully controlled. For recognition "in the wild", large variances in face pose and illumination, cluttered backgrounds, occlusions, audio and video noise, as well as issues with subtle cues of expression are some of the issues to target. In this paper, we describe a multimodal approach for video-based emotion recognition in the wild. We propose using summarizing functionals of complementary visual descriptors for video modeling. These features include deep convolutional neural network (CNN) based features obtained via transfer learning, for which we illustrate the importance of flexible registration and fine-tuning. Our approach combines audio and visual features with least squares regression based classifiers and weighted score level fusion. We report state-of-the-art results on the EmotiW Challenge for "in the wild" facial expression recognition. Our approach scales to other problems, and ranked top in the ChaLearn-LAP First Impressions Challenge 2016 from video clips collected in the wild.

1. Introduction

Audio- and video-based emotion recognition in the wild is challenging, because of noise, large idiosyncratic variances, and sensor-related differences. This paper describes a multimodal approach for audio-visual emotional expression recognition. Our approach processes complementary features with summarizing functionals, classifies these descriptors with a set of least squares based classifiers, and combines their output with decision level fusion.

The Emotion Recognition in the Wild (EmotiW) Challenge provides out of laboratory data – Acted Facial Expressions in the Wild (AFEW) – collected from videos that mimic real life, and poses a very difficult and realistic problem [5–7].

This paper extends our contribution to the EmotiW 2015 Challenge [22], which was ranked second in the official competition, in terms of the introduced approach, level of detail, and amount of experimental validation. Ref. [22] proposed the combination of multiple visual features with audio over summarizing functionals. Here, we extend this framework by employing deep convolutional neural network (CNN) features, as well as an investigation of suitable transfer learning strategies that can be used with CNNs. We show that incorporating multiple registration schemes into the model helps transfer learning. We also illustrate the success of the approach on two additional in the wild datasets, namely, the ChaLearn-LAP 2016 First Impressions Challenge database, and the EmotiW 2016 corpus, which is an extension of the 2015 corpus. Our system achieves the best official result in the ChaLearn-LAP First Impressions Challenge. We further validate our visual features using two widely used lab-controlled corpora, namely the Extended Cohn–Kanade dataset (CK+) [33] and MMI [50].

The remainder of this paper is organized as follows. In the next section we discuss related work on video-based facial expression recognition in naturalistic settings. Section 3 describes the proposed approach. In Section 4 we briefly introduce the corpora and baseline feature sets. In Section 5, we give experimental results. Finally, Section 6 concludes the paper and summarizes our findings.

2. Related work

Facial affective displays in real life involve subtle changes, in contrast to the exaggerated displays in posed expressions [60]. Therefore, facial expression recognition "in the wild" poses much greater challenges, and the recognition accuracies are invariably

☆ This paper has been recommended for acceptance by Mohammad Soleymani.
* Corresponding author.
E-mail addresses: hkaya@nku.edu.tr (H. Kaya), furkan.gurpinar@boun.edu.tr (F. Gürpınar), salah@boun.edu.tr (A.A. Salah).

http://dx.doi.org/10.1016/j.imavis.2017.01.012
0262-8856/

much lower. Furthermore, controlled illumination and pose settings in the lab are not available in the wild, which adds purely physical challenges to the problem, and causes issues starting from the detection phase [59].

We focus here only on video-based approaches, which implies both dynamic and multimodal information. Multimodal approaches to emotional expression recognition leverage both paralinguistic audio cues, as well as the synchronization between modalities to improve robustness. Early multimodal approaches focused on coarse affective states (e.g. positive and negative states), because data collection and annotation for natural or spontaneous scenarios were difficult (see Ref. [60] for a comprehensive survey of earlier approaches, and Ref. [55] for available databases).

Together with developments of multimodal expression research in more natural conditions, it became obvious that the temporal dynamics of expressions contained rich information [61]. Krumhuber et al. previously used a 2-person trust game setting to illustrate that facial dynamics significantly affect social judgments [27]. Research on smile dynamics illustrated fake and genuine enjoyment smiles have quite distinctive dynamics, and that humans are sensitive to such cues [9]. However, automatic classification of subtle distinctions requires controlled and high-quality data, which cannot be assumed for naturalistic settings. The Emotion Recognition in the Wild (EmotiW) Challenge, which started in 2013, initiated an effort to overcome challenges of data collection, annotation, and standardized testing for multimodal and dynamic emotion recognition in the wild. The challenge used the AFEW corpus, which was collected from movies with close-to-real-world conditions [7]. It was quickly understood that for image-based processing, it was particularly important to get good face detection and alignment, as well as rejection of non-face images [29]. In the top performing system of the first challenge, visual bag of words features, gist features, paralinguistic audio features and Local Phase Quantization from Three Orthogonal Planes (LPQ-TOP) features obtained from aligned faces were each processed by RBF kernels, and combined with a multi kernel support vector machine [45]. The contribution of audio was very small (its normalized weight was 1.8, as opposed to 50.1 for visual HOG features). The weight for features extracted over the entire scene was even smaller than the audio.

In the 2013 EmotiW Challenge, Liu et al. treated images from each video clip as a set and represented them as points in a Grassmanian manifold [29]. Audio was separately modeled and fused linearly with video-based classifiers. One year later, they extended their approach by representing each video clip using three kinds of image set models (i.e. linear subspace, covariance matrix, and Gaussian distribution) respectively. As features, they used HOG, dense SIFT, as well as features extracted from the 9-layer deep convolutional neural network (CNN) pre-trained with ImageNet [26]. This system achieved the best test set accuracy of 50.37%, and represented a significant improvement over the systems submitted to the 2013 challenge. The first runner-up system proposed a hierarchical voting classifier, and showed that the addition of audio features had a small, but persistent impact (from 44.72% to 47.17%) [47]. Multiple kernel learning and SVM were popular in the submissions, but the top system used a partial least squares based classifier.

In both years, surprise and disgust were the most difficult classes to recognize, whereas happy, angry, and neutral had relatively high accuracies. By 2015, it was clear that deep neural networks harbored great potential for describing features for non-controlled settings. Their main advantage was a resistance to alignment issues, as well as to noise. In their submission to the 2015 EmotiW Challenge, Ebrahimi Kahou et al. used a Recurrent Neural Network (RNN) combined with a CNN to model the expression dynamics [11]. Their results suggested that temporal integration was better than averaging per-frame decisions. They also combined feature level and decision level fusion, noting that such powerful and complex models were prone to overfitting the training set, as evidenced by the great discrepancy between training and validation set accuracies (98.3% vs. 26.6% for activity modality). Their solution was to adopt early stopping during training. The usage of rectified linear units and dropout strategies is also commonly employed to control overlearning in CNNs [26]. These approaches will directly benefit from increases in the amount of training data. The top performing system used a CNN model as well, but it fused the CNN model output with one audio and three linear SVMs trained with AU-aware facial feature relations on two face scales [56].

The number of examples in the challenge database is increased to 1645 videos (AFEW 5.0) during the years, but the database still has some shortcomings. Disgust, sadness and surprise were found to be really difficult to reliably classify on this database. There are few other initiatives to collect and evaluate in the wild data. The AM-FED corpus contains 242 videos of people watching commercials on a computer [35]. The RECOLA database presents 27 videos that are continuously annotated in time and space for arousal and valence dimensions [42]. A very recent database is collected by Zafeiriou et al. from YouTube, containing around 500 videos and annotated with regard to valence and arousal [58].

3. Proposed method

The proposed approach is illustrated in Fig. 1, and contains the detection of the face, its alignment (registration) with a fixed model or with a set of facial landmarks, and subsequent feature extraction and classification. We use two separate alignment options, which greatly improves processing in the CNN. The CNN pre-training and fine tuning stages are carried out prior to processing of the target emotional video corpus. Audio is separately processed and fused at the decision level. We describe each phase in the pipeline separately.

3.1. Preprocessing and image purification

Facial registration is one of the most important steps in face image processing. To guard against registration errors, we applied a purification step in processing faces of the EmotiW dataset. To deal with rotated faces and false positives in face detection, we use a principal component analysis (PCA) based method to automatically remove false detections, as shown to be effective in Refs. [24,29,47]. The idea is to measure the mean reconstruction error per image, after projecting the images to the PCA space and back. We discard the frames with a high reconstruction error, as these are probably poorly detected or poorly aligned images. We remove videos that have less than three valid images from the validation set. For the sequestered test set, all instances are retained. After purification, images are resized to 64×64 pixels. We manually remove poorly aligned images from the training set, which improves the quality of training.

3.2. Visual descriptors

We extract and compare Scale Invariant Feature Transform (SIFT) [32], Histogram of Oriented Gradients (HOG) [3], Local Phase Quantization (LPQ) [17,19], Local Binary Patterns (LBP) [37] and its Gabor extension (LGBP), as well as deep convolutional neural network (CNN) based visual descriptors. For LPQ, LBP, and LGBP, the Three Orthogonal Planes (TOP) extension is popularly used in video modeling [62]. This extension applies the relevant descriptor on XY, XT and YT planes (where T represents time) independently, and concatenates the resulting histograms. Also in our implementation, we divide the video into two equal length volumes over the time axis and extract spatio-temporal TOP features from each volume to further enhance temporal modeling. In the following, we provide brief explanations of LPQ, LBP and LGBP descriptors.